Interspeech 2018
2-6 September 2018, Hyderabad

Real-Time Scoring of an Oral Reading Assessment on Mobile Devices

Jian Cheng

Analytic Measures Inc., Palo Alto, California, U.S.A.

Jian.Cheng@AnalyticMeasures.com

Abstract

We discuss the real-time scoring logic for a self-administered oral reading assessment on mobile devices (Moby.Read) to measure the three components of children's oral reading fluency skills: words correct per minute, expression, and comprehension. Critical techniques that make the assessment real-time on-device are discussed in detail. We propose the idea of producing comprehension scores by measuring the semantic similarity between the prompt and the retelling response utilizing the recent advance of document embeddings in natural language processing. By combining features derived from word embedding with the normalized number of common types, we achieved a human-machine correlation coefficient of 0.90 at the participant level for comprehension scores, which was better than the human inter-rater correlation 0.88. We achieved a better human-machine correlation coefficient than that of the human inter-rater in expression scores too. Experimental results demonstrate that Moby.Read can provide highly accurate words correct per minute, expression and comprehension scores in real-time, and validate the use of machine scoring methods to automatically measure oral reading fluency skills.

Index Terms: assessment, oral reading fluency, literacy, expression, comprehension

1. Introduction

Moby.Read is a new, self-administered, fully automated oral reading fluency assessment developed for K-5 students [1, 2, 3]. The prototype system was built on an iPad mini 4 as a standalone app. In each test session, students are asked to read a word list, read an easy practice passage, and read three grade-level passages. After reading one passage aloud, students are asked to retell the passage in their own words, put in all the details they can remember, then answer two short questions aloud.

Fluency is the ability to "read text with speed, accuracy, and proper expression" [4]. In this paper, we focused our automatic scoring logic on passage reading (PRead) to produce Words Correct Per Minute (WCPM) and reading expression scores, on passage retelling (PRetell) to produce reading comprehension scores. WCPM is a score based on the number of words read correctly in a minute of reading, an informative measure of oral reading fluency [5]. Expression is the degree that a student can clearly express the meaning and structure of the text through appropriate intonation, rhythm, phrasing, and emphasis that will enhance understanding and enjoyment in a listener [6]. Comprehension is the degree that a student can retell major and minor concepts/themes/facts in the original passage. Scoring of expression and comprehension will emphasize reading for meaning instead of reading for speed. Automatic scoring can reduce the need for teacher training and help ensure consistency.

Scores are produced in real-time on-device. The advantages of real-time on-device are that we may provide scores and feedback immediately; we may select appropriate reading materials adaptively based on the past or real-time performance of a specific student to level the student more accurately, etc.

2. The mobile speech recognition system

Although automatic speech recognition (ASR) is a compute-intensive process, feasible, on-device speech recognition was researched [7, 8, 9, 10, 11]. With the recent introduction of the neural processing unit (NPU) or neural engine to mobile devices, we expect that complex acoustic models and language models can be implemented on the latest mobile devices to achieve better performance. Speech recognition on-device will not be a barrier for a wide class of mobile applications. In the following subsections, we introduce our system.

2.1. Acoustic models

The acoustic model used for speech recognition on-device is a Deep Neural Network - Hidden Markov Model (DNN-HMM) [12] that contains 4 hidden layers and 300 p-norm (p = 2) non-linearity neurons with a group size G = 10 [12] per hidden layer, trained using all of Librispeech's training sets [13]: 960 hours of clean native (L1) reading data. The sample rate for the on-device speech recognition is 8,000. The inputs of the DNN are 40-dimensional log mel-filterbank energies calculated on a 25ms window every 10ms, and the output dimension is 2,064 context-dependent triphone states. Both left and right context are 6.

There are several model mismatch issues that may degrade the ASR performance: 1) an adult acoustic model was used to recognize children speech; 2) narrowband was used; 3) the acoustic model was trained using very clean/quiet recordings, so the ASR accuracy may diminish with very noisy data. Despite these issues, the overall on-device acoustic model performance is good since we deal with very low perplexity situations with suitable language models.

In our previous work [14, pp. 24], the model mismatch effect was researched, such as checking the child test set performance when adult acoustic models are used. We concluded that DNNs seem to be good at learning invariant representations of speech signals, and adult data could be more suitable for learning speech representations. When using mismatch adult acoustic models, the performance damage to constrained item types is not so severe. Still, after we collect enough child responses and transcriptions, we plan to train a better acoustic model by combining Librispeech's data with children's speech data. Domain-specific training data always help [14].

2.2. Language models

Item-specific rule-based language models (RBLMs) [15, 16] are built for PRead. No data from this study was used to tune the RBLMs. Item-specific 3-gram language models were built for PRetell, using all the human transcriptions we have for the specified item, around 59 transcriptions per item with the averaged

10.21437/Interspeech.2018-34

vocabulary size 230. The comprehension scores reported in this paper are biased by the fact that we used the same data to train the language models. We assume that this bias is not big since we deal with a very narrow domain.

2.2.1. Advantages of item-specific language models

With the constraints that the decoding should be finished in real-time and the decoding devices are mobile, the ASR accuracy decreases significantly when the language model is big. We deal with a narrow domain with expected answers. Building a small and constraint language model can help to decode quickly and achieve good accuracy. It helps to overcome the challenge that children's speech contains larger acoustic variability because of their variable vocal tract length and formant frequencies [17] and some other model mismatch challenges. It gives the benefit of the doubt for accent or non-native speech. Many spoken assessment applications fall to the category that has a narrow domain with expected answers. Item-specific language models with smaller vocabulary sizes are preferred, and are often used in practice for spoken assessment applications [14].

2.2.2. The rule-based language model

For each expected passage, sentence, phrase or token sequence, a simple direct graph is built that has a path from the first word in the sequence to the last word [15, 16]. Different direct arcs with probabilities are added to represent different classes of changes made by subjects, such as skipping, repeating, inserting, and substituting. Adding a back-off arc will allow domain words to be spoken in any order. Both changes and probabilities can be learned from data. Using domain data can help to build better language models. The graphs generated from several different expected answers can be combined together with the expected probabilities as the final RBLM. Naturally, RBLMs give the expected answer sequences higher probabilities, the less likely orders lower probabilities. RBLMs can be compiled on-line. It gives us the flexibility to recognize any contents that are generated dynamically. Humans can add arbitrary reasonable rules to be used by RBLMs directly.

2.3. The ASR decoder

The decoding engine [18] is based on KALDI [19]. The modifications were made to utilize mobile devices' single instruction multiple data (SIMD) and digital signal processor (DSP) frameworks. The supporting utils were built to convert RBLMs to finite state transducers (FSTs) for decoding. When we start to record responses, the engine decodes progressively every 0.128 seconds. The decoding real-time factor floats around 0.2 on an iPad Mini 4. We chose the acoustic scale so that insertions and deletions are balanced to avoid ignoring the speech signal.

2.4. An ASR performance comparison

Although Google cloud speech API [20] (GSpeech) can be used off-the-shelf without any additional modifications, the word error rates (WERs) are rather high on our children reading and retelling tasks. The main reason is that GSpeech is designed for recognizing any general English with a broad language model. Its purpose is too general to perform well in this narrow domain. For the 282 PRead responses, GSpeech achieved WER 34.8% (n=23,736) and our on-device ASR engine achieved WER 10.7%. For the 282 PRetell responses, GSpeech achieved WER 32.9% (n=11,070) and our on-device ASR engine achieved WER 16.3%. Our server-side ASR engine that used broadband speech, featured more elaborate acoustic models, and utilized a larger beam value can achieve WER 9.6% for the 282 PRetell responses.

3. Machine scoring methods

3.1. Word correct per minute

The number of correct words is derived by using the ASR result to do edit-distance with the prompt. The insertions caused by disfluencies were ignored. The time duration is from the beginning of the first correct word to the end of the last correct word according to the ASR result. The session level WCPM is the median WCPM value from the three passages, a widely used procedure in measuring oral reading fluency [21].

3.2. Expression

Although our engine generates a lot of features, only relevant and normalized ones were used (Table 1). These features don't depend on the length of the materials produced. The difference between log_seg_prob and $nlog\_seg\_prob$, $iw\_log\_seg\_prob$ and $niw\_log\_seg\_prob$ is that for the latter: a) when we built native duration statistics, the durations were normalized by the articulation rate of each response; b) when we computed segmental probabilities, the durations of segments were normalized by the articulation rate of the corresponding response. These can help remove the effects of the speaking rate, which usually has a strong correlation with human expression scores.

Table 1: *The features used to predict expression scores.*

feature	description
log_leading_sil	Log the leading silence duration.
art	Articulation rate: the number of phonemes per second of speech.
ros	Rate of speech: the number of phonemes per second of speech and inter word pauses.
log_seg_prob	The averaged log likelihood segmental probability for phonemes [22] based on Librispeech native statistics.
iw_log_seg_prob	The averaged log likelihood segmental probability for inter word silences [22] based on Librispeech native statistics.
nlog_seg_prob	The normalized version of log_seg_prob.
niw_log_seg_prob	The normalized version of iw_log_seg_prob.
amloglike	The acoustic model log likelihood of the recognized result normalized by the total number of frames.
lmloglike	The language model log likelihood of the recognized result normalized by the total number of frames.

3.3. Comprehension

In natural language processing (NLP), it is becoming popular to use neural network based unsupervised learning algorithms to represent variable-length pieces of texts, such as words, sentences, passage, and documents as fixed-length real value feature representations that encode the meaning of texts. For these methods (e.g. word2vec [23] or doc2vec [24]), the training objective is usually to learn better word/document vector representations so that they can be used to predict the nearby words with higher probabilities. As a consequence, in the trained continuous vector space semantically similar words or documents are mapped to similar positions. Meaningful results (e.g. *king - man + women = queen*) can be obtained by adding/subtracting these vectors. These methods achieved better performance in many NLP tasks [23, 24].

We seek semantic similarity measurements between the prompt passages and the retelling responses that are automated and objective. The vector representations of documents could

be a good fit. Comparing to previous works [25, 26], our task could be easier to handle since the domain has been constrained by the prompts.

The number of words spoken or the number of different words used could be a good indicator of the similarity if the subjects are in good-faith, although such nonlinguistic surface features are too superficial. We are more interested in the measurements that can check semantic similarity directly, and don't have strong correlations with these surface features. The number of common spoken tokens or the number of common spoken types could be a good semantic similarity indicator, but it could fail when a subject uses semantic similarity words or phases that are not in the prompt. Using word2vec or doc2vec may fix some issues. Assume that every response can be converted to a vector that represent the whole content of the response, the semantic similarity between two documents may be computed by checking the distance or similarity of two vectors. As a result, we proposed features w2v_ed, d2v_cos, LSI_cos. All potentially useful similarity metrics for the comprehension scores we are interested in are listed in Table 2.

As words can be represented by real number vectors, we may use the centroid of the word vectors of the text to represent the text. Usually it makes sense to remove stop words before computing the centroid. We did observe the performance gain by doing so. We can use either cosine similarity or Euclidean distance between two vectors to serve as a measure of the similarity between two texts. For w2v, we observed a significant performance gain by using Euclidean distance. We used the simple average of the word vectors of the text as the centroid to represent the text. We didn't observe any performance gain when using TF-IDF weighted average of word vectors. Furthermore, we may use different statistical functions to aggregate the word vectors to represent the document. We concatenated 4 statistical vectors (mean, minimum, maximum, media) together to form a 4 * 300 = 1200 dimension vector for a document. It can produce better results. We hypothesize that the distribution of word embedding vectors plays an important role to represent the document. The statistical vectors may catch some properties of the distribution.

We used Googles word2vec pre-trained vectors that were trained on part of Google News dataset (about 100 billion words). The archive is available online as GoogleNews-vectors-negative300.bin.gz [27]. The model contains 300-dimensional vectors for 3 million words and phrases.

Table 2: *Some potential useful similarity metrics between prompts and responses for comprehension scores.*

feature	description
ntokens_n	The number of words were spoken in the response normalized by the number of words in the prompt.
ntypes_n	The number of different words were spoken in the response normalized by the number of different words in the prompt. It is a measure of the vocabulary size in the response.
nctypes_n	The number of the same words between the prompt and the response normalized by the number of different words in the prompt. It is a measure of the overlapped vocabulary size between the prompt and the response.
w2v_ed	Euclidean distance between two documents' statistical vector representations that are derived from word embeddings after removing stop words.
d2v_cos	Cosine similarity between two documents' vector representations based on doc2vec [24].
wmd	Word mover's distance [28] between two documents based on word embeddings after removing stop words.
LSI_cos	Cosine similarity between two documents' vector representations derived from Latent Semantic Indexing based on the term vector model [29].

4. Experimental results

A preliminary study of the Moby.Read system was conducted with a sample of 99 children in grades 2-4 from four different elementary schools [1, 2]. Recordings of PRead and PRetell of three grade-level unpracticed passages from the preliminary study were used. 5 children who barely produced meaningful responses (silence or inaudible) were excluded from this study. The total number of subjects in this study are 94. The details about the raters' qualifications, training procedures, rubrics and how to derive human WCPM, expression and comprehension scores can be found in [2].

The results reported were produced on a simulation platform that mimicked the conditions as if it were run on mobile devices. The real-time turnaround was verified on mobile devices. The system was published on Apple App Store [3].

4.1. Word correct per minute scores

A scatter plot of WCPMs between human and machine was shown in Figure 1. The correlation between two expert raters is 0.991. It repeated the conclusion we knew: machine can produce reliable WCPMs for oral reading fluency [16, 30, 31].

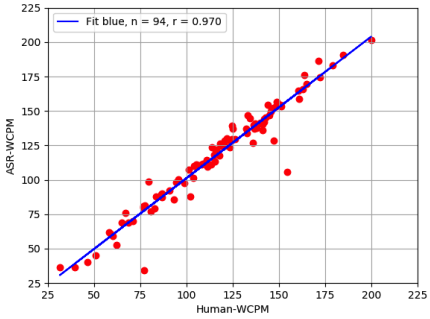

Figure 1: *Session-level scatter of median WCPM.*

We listened to the 6 recordings of two outliers. Both spoke quite softly and one mumbled the readings for certain time periods. The consequence of low Signal-to-Noise Ratio (SNR) makes it difficult to recognize certain part of signals. Google cloud speech API only recognized 25.2% words correctly and ignored most of the signal for these 6 recordings. The same kind of outliers and issues for kids were identified before [32]. Addressing low SNR effectively by instructions, e.g. avoiding high background noise and low speech volume, is the key solution we are looking for. "Be in a QUIET place" is on the sign-in page of the app. Speaking clearly instead of mumbling so that others can hear is the requirement. Reliable scores depend on audible speech.

4.2. Oral reading expression scores

Every recording of PRead was rated by 3 different human raters for their 'Oral Reading Expression' scores on 6 categories, with 5 representing the best rating and 0 representing silence or irrelevant or completely unintelligible material. The rating distribution is: 0, n=5; 1, n=37; 2, n=95; 3, n=263; 4, n=304; 5, n=142. The average of the correlations of human raters who correlate with the average of others at the response level is 0.795. For the 3 pairs of raters, the average of the inter-rater correlations at the response level was 0.740.

The correlations among features we discussed in Subsection 3.2 and human ratings at the response level are shown in Table 3. The speech rate features have the highest correlations

with human ratings. Putting more weights on the features normalized by speaking rate may downplay the role of rate.

Table 3: *Feature cross correlations for expression scores.*

	hv	1	2	3	4	5	6	7	8
1:log_leading_sil	-0.18								
2:ros	0.82	-0.22							
3:art	0.77	-0.20	0.92						
4:log_seg_prob	0.81	-0.19	0.90	0.92					
5:iw_log_seg_prob	0.50	-0.17	0.62	0.38	0.45				
6:nlog_seg_prob	0.60	-0.11	0.65	0.69	0.65	0.40			
7:niw_log_seg_prob	0.31	-0.12	0.39	0.14	0.23	0.87	0.27		
8:amloglike	0.59	-0.10	0.61	0.49	0.58	0.40	0.60	0.28	
9:lmloglike	-0.57	0.25	-0.82	-0.69	-0.68	-0.70	-0.62	-0.53	-0.55

The final session-level expression is an average of individual expression scores. If a response doesn't have enough information to generate an expression score, it will be ignored when computing the final session-level expression score.

Using a neural network model 10-fold cross-validation, we achieved correlation 0.856 (0.902) in response (session) level. This is better than a linear regression model 0.840 (0.887). We made sure different folds have no overlap of the same subjects.

4.3. Comprehension scores

Every recording of PRetell was rated by at least 4 different human raters for their 'Retelling Comprehension' scores on 7 categories, with 6 representing the best rating and 0 representing silence or irrelevant or completely unintelligible material. On average, there are 4.5 ratings per response. The rating distribution is: 0, n=25; 1, n=128; 2, n=173; 3, n=288; 4, n=266; 5, n=203; 6, n=179. The average of the correlations of human raters who correlate with the average of others at the response level is 0.842. For the 11 pairs of raters who have more than 100 common ratings, the average of the inter-rater correlations at the response level was 0.786.

Table 4: *Feature cross correlations for comprehension scores.*

	hv	1	2	3	4	5	6
1:ntokens_n	0.82						
2:ntypes_n	0.84	0.95					
3:nctypes_n	0.87	0.85	0.90				
4:w2v_ed	-0.83	-0.78	-0.84	-0.84			
5:wmd	-0.85	-0.75	-0.79	-0.92	0.88		
6:d2v_cos	0.71	0.61	0.60	0.73	-0.68	-0.79	
7:LSI_cos	0.65	0.52	0.54	0.72	-0.65	-0.83	0.83

We calculated the cross correlations (Table 4) at the response level among features we discussed in Subsection 3.3 and human ratings. These features were extracted using human transcriptions. The performances of d2v_cos and LSI_cos depend on the training settings: e.g. the training corpus, random seeds and setting parameters. In Table 4 we reported the best results we achieved for d2v_cos and LSI_cos from several trials. Because of the limited domain data, the potential overfitting and weaker correlations comparing to others, we didn't explore them further. All other results didn't involve overfitting.

It can be seen that the normalized number of different words spoken in the response is a good indicator of comprehension. When the subject is in the good-faith (it is almost always the case for K-5 grade kids), it makes sense since comprehension will depend on the complexity of the materials produced. By the nature of PRetell, a lot of term overlap is expected. The table reflects that only considering the words in the prompt can improve the performance significantly. Among the features that utilize the word embedding similarities by considering and weighting the semantically similar words that are not in the prompt and are ignored by nctypes, w2v_ed and wmd are good ones.

There is no setting parameter for the features nctypes_n, w2v_ed, wmd. We noticed that wmd has a strong correlation with nctypes_n. At the same time, wmd used the same word embedding as w2v_ed. In that sense, w2v_ed could be better to enhance the final performance. We scaled nctypes_n, w2v_ed to the range [0, 1] and then used their simple average as our final metric and achieved r=0.888 at the response level.

All the comprehension performance discussed so far is based on human transcriptions. In the real application, we used the ASR recognition results to do the computation. It drags down the performance a little bit. Following the same procedures discussed but using the ASR transcriptions, we achieved r=0.903 at the session level (Figure 2). After collecting enough data, using a complex supervised machine learning model that can combine different features discussed in Table 2 together may further improve the final performance.

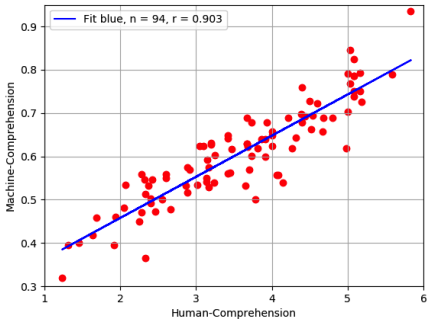

Figure 2: *Session-level scatter of comprehension.*

5. Conclusions

We built an oral reading assessment system on mobile devices that delivers reliable WCPM, expression, and comprehension scores in real-time for first-language learners in grades 2-4 [3]. Our RBLMs relieve the requirement of field data collection for new reading passages to produce WCPM and expression scores; however, data collection is still required for passage retellings in order to build suitable language models to achieve the best performance. The proposed idea of producing comprehension scores by measuring the semantic similarity between the prompt passage and the retelling response utilizing the document embeddings works well. For both expression and comprehension scores, the human-machine correlations are better than the human inter-rater ones, which validates the effectiveness of the system. The findings support the use of machine scoring methods to measure oral reading fluency skills automatically.

We expect the system can be highly useful beyond the application discussed here, such as in second-language learning for adults as well as children. Assessing in real-time means the system can rapidly adapt to a learner's performance, which can be used by learning systems to condition immediate, personalized feedback and select the next challenge within a session.

6. Acknowledgements

The research described here was supported by the Institute of Education Sciences, U.S. Department of Education, through the Small Business Innovation Research (SBIR) program contract ED-IES-16-C-0004 and ED-IES-17-C-0030 to Analytic Measures Inc. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

7. References

[1] J. Bernstein, J. Cheng, J. Balogh, and E. Rosenfeld, "Studies of a self-administered oral reading assessment," in *SLaTE*, 2017.

[2] J. Bernstein, J. Cheng, J. Balogh, and R. Downey, "Artificial intelligence for scoring oral reading fluency," in *Applications of artificial intelligence to assessment*, H. Jiao and R. W. Lissitz, Eds. Charlotte, NC: Information Age Publisher, 2018, forthcoming.

[3] Analytic Measures Inc., "Moby.Read Lite," 2018, Apple App Store. [Online]. Available: https://itunes.apple.com/us/app/moby-read-lite/id1292957097?mt=8

[4] National Institute of Child Health and Human Development, "Report of the national reading panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," U.S. Government Printing Office, Tech. Rep. NIH Publication No. 00-4769, 2000.

[5] M. M. Wayman, T. Wallace, H. I. Wiley, R. Tich, and C. A. Espin, "Literature synthesis on curriculum-based measurement in reading," *Journal of Special Education*, vol. 41, no. 2, pp. 85–120, 2007.

[6] J. Miller and P. J. Schwanenflugel, "A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children," *Reading research quarterly*, vol. 43, no. 4, pp. 336–354, 2008.

[7] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *ICASSP*, 2006, pp. 185–188.

[8] B. Zhou, X. Cui, S. Huang, M. Cmejrek, W. Zhang, J. Xue, J. Cui, B. Xiang, G. Daggett, U. V. Chaudhari, S. Maskey, and E. Marcheret, "The IBM speech-to-speech translation system for smartphone: improvements for resource-constrained tasks," *Computer Speech and Language*, vol. 2, pp. 592–618, 2013.

[9] X. Lei, A. Senior, A. Gruenstein, and J. Sorensen, "Accurate and compact large vocabulary speech recognition on mobile devices," in *Interspeech*, 2013, pp. 662–665.

[10] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, "Personalized speech recognition on mobile devices," in *ICASSP*, 2016, pp. 5955–5959.

[11] CMUSphinx, "Pocketsphinx," Computer Software. GitHub. [Online]. Available: https://github.com/cmusphinx/pocketsphinx

[12] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *ICASSP*, 2014, pp. 215–219.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[14] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Commun.*, vol. 73, no. C, pp. 14–27, Oct. 2015.

[15] J. Cheng and B. Townshend, "A rule-based language model for reading recognition," in *SLaTE*, 2009.

[16] J. Cheng and J. Shen, "Towards accurate recognition for children's oral reading fluency," in *IEEE-SLT*, 2010, pp. 91–96.

[17] H. Vorperian and R. Kent, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," *Journal of Speech, Language, and hearing research*, vol. 50, no. 6, pp. 1510–1545, 2007.

[18] Analytic Measures Inc., "amiASR," 2018, Computer Software. Apple App Store, Version 1.0. [Online]. Available: https://itunes.apple.com/us/app/amiasr/id1032618938?mt=8

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The KALDI speech recognition toolkit," in *IEEE-ASRU*, 2011.

[20] Google, "Google cloud speech API." [Online]. Available: https://cloud.google.com/speech/

[21] Y. Petscher and Y.-S. Kim, "The utility and accuracy of oral reading fluency score types in predicting reading comprehension," *Journal of school psychology*, vol. 49, no. 1, pp. 107–129, 2011.

[22] J. Cheng, "Automatic assessment of prosody in high-stakes English tests," in *Interspeech*, 2011, pp. 1589–1592.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[24] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *ICML*, 2014, pp. 1188–1196.

[25] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre, "SemEval-2012 Task 6: a pilot on semantic textual similarity," in *SemEval*, 2012, pp. 385–393.

[26] E. Agirre, C. Banea, D. M. Cer, M. T. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2016 Task 1: semantic textual similarity, monolingual and cross-lingual evaluation," in *SemEval*, 2016, pp. 497–511.

[27] Google, "word2vec," Computer Data. [Online]. Available: https://code.google.com/archive/p/word2vec/

[28] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *ICML*, 2015, pp. 957–966.

[29] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[30] J. Balogh, J. Bernstein, J. Cheng, A. Van Moere, B. Townshend, and M. Suzuki, "Validation of automated scoring of oral reading," *Educational and Psychological Measurement*, vol. 72, no. 3, pp. 435–452, 2012.

[31] D. Bolanos, R. A. Cole, W. H. Ward, G. A. Tindal, J. Hasbrouck, and P. J. Schwanenflugel, "Human and automated assessment of oral reading fluency," *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1142–1151, 2013.

[32] J. Cheng, Y. Zhao D'Antilio, X. Chen, and J. Bernstein, "Automatic spoken assessment of young English language learners," in *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014, pp. 12–21.